CLOUDFLARE®

RPKI: our approach for deploying at scale

Louis Poinsignon - RIPE77

Introduction to Cloudflare

Some numbers...

- 155+ PoPs and growing
- 72+ countries
- 186+ Internet exchanges

- >600B Web requests a day ~10% of all web requests
- Regular DDoS attacks larger than 500Gbps, 300M PPS
- >100bn DNS requests a day

CLOUDFLARE

Who am I?

- Louis Poinsignon
- Network, data and software @ Cloudflare London and SF
- Built a network data pipeline (flows and BGP) for Cloudflare scale, open-source:
  https://github.com/cloudflare/goflow
  https://github.com/cloudflare/fgbgp

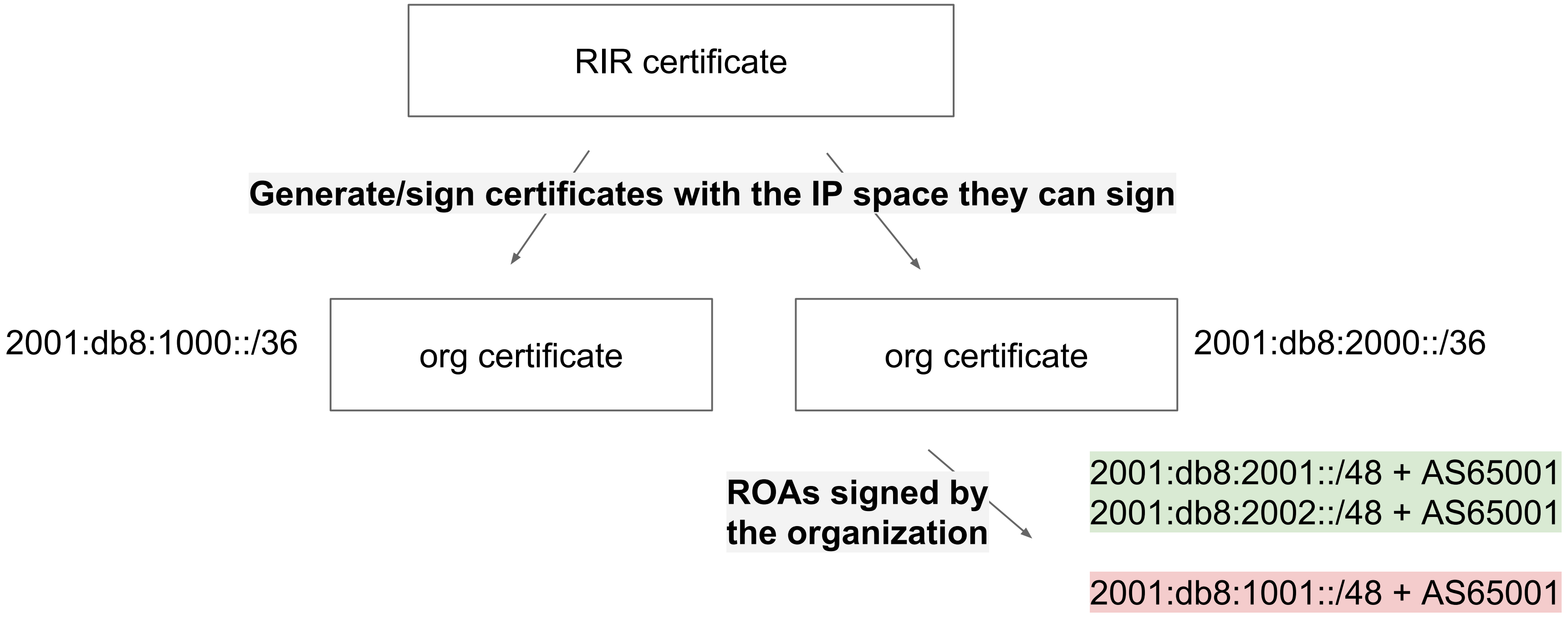
RPKI

What is RPKI

RFC6480: defines a way of cryptographically signing:
route + length + origin ASN

RFC6810: defines a communication method between router and validation system

How RPKI works



RIR certificate

Generate/sign certificates with the IP space they can sign

2001:db8:1000::/36

org certificate

org certificate

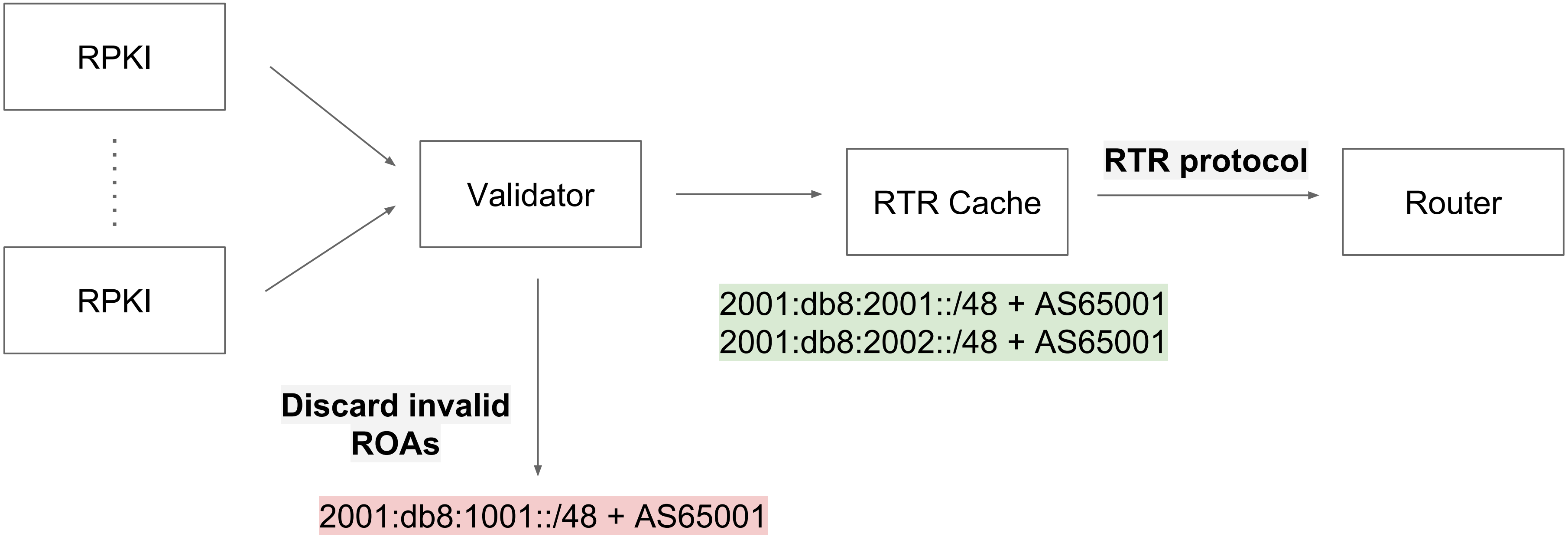
2001:db8:2000::/36

ROAs signed by the organization

2001:db8:2001::/48 + AS65001
2001:db8:2002::/48 + AS65001

2001:db8:1001::/48 + AS65001

How validation works



RPKI

RPKI

Validator

RTR Cache

RTR protocol

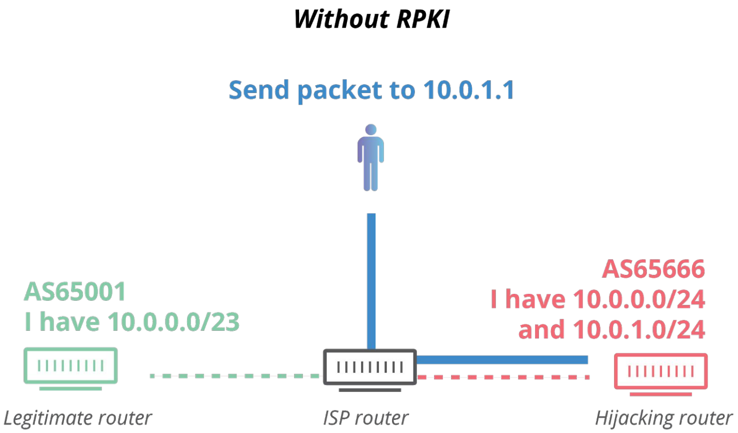
Router

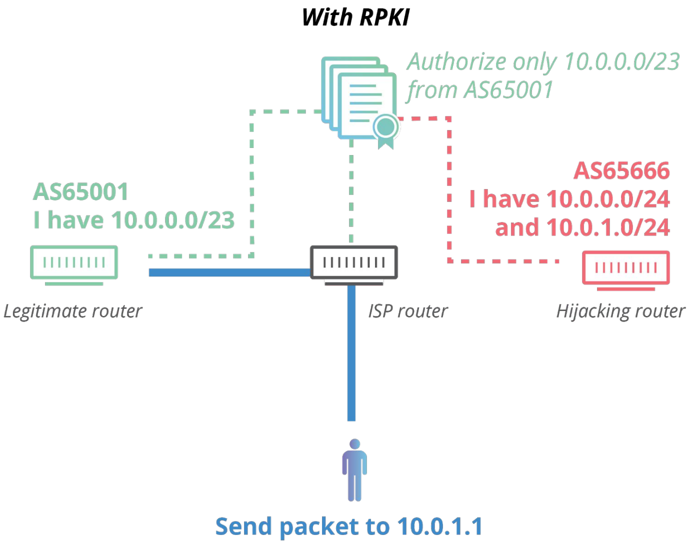
2001:db8:2001::/48 + AS65001
2001:db8:2002::/48 + AS65001

Discard invalid ROAs

2001:db8:1001::/48 + AS65001

Summary


Without RPKI
Send packet to 10.0.1.1
AS65001
I have 10.0.0.0/23
AS65666
I have 10.0.0.0/24
and 10.0.1.0/24
Legitimate router
ISP router
Hijacking router

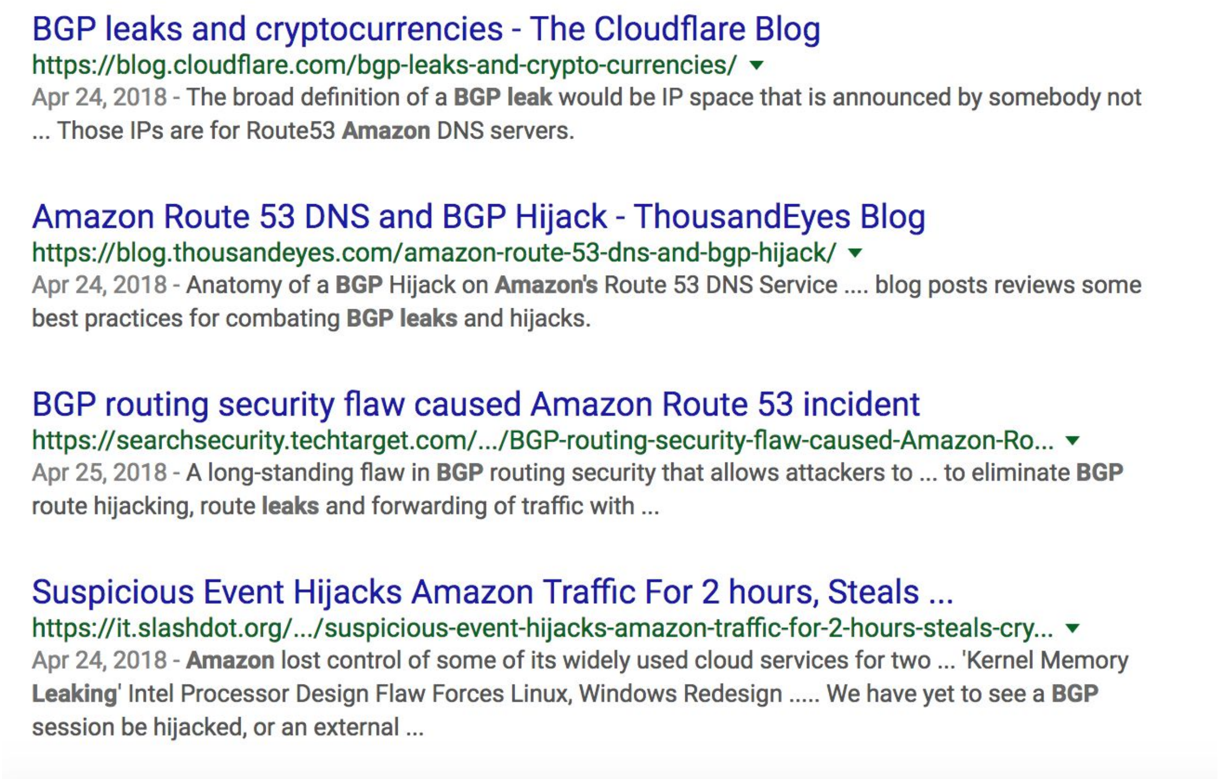

With RPKI
Authorize only 10.0.0.0/23
from AS65001
AS65001
I have 10.0.0.0/23
AS65666
I have 10.0.0.0/24
and 10.0.1.0/24
Legitimate router
ISP router
Hijacking router
Send packet to 10.0.1.1

Use-cases

- Filter out bad announcements
- For “Bring your own IP” services → make sure your clients are the true owner of a range

BGP leaks and hijacks

Why signing?



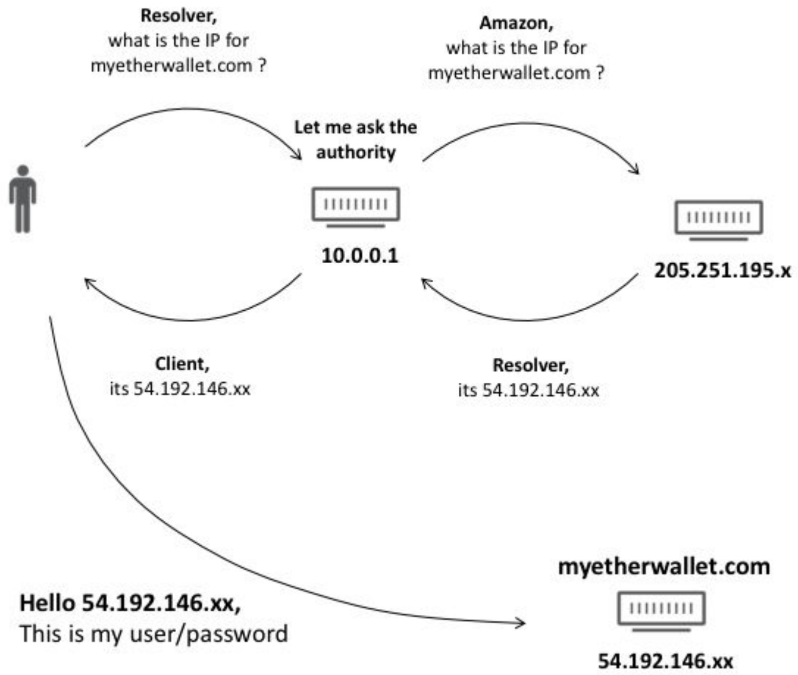
BGP leaks and cryptocurrencies - The Cloudflare Blog
https://blog.cloudflare.com/bgp-leaks-and-crypto-currencies/
Apr 24, 2018 - The broad definition of a **BGP leak** would be IP space that is announced by somebody not ... Those IPs are for Route53 **Amazon** DNS servers.

Amazon Route 53 DNS and BGP Hijack - ThousandEyes Blog
https://blog.thousandeyes.com/amazon-route-53-dns-and-bgp-hijack/
Apr 24, 2018 - Anatomy of a **BGP** Hijack on **Amazon's** Route 53 DNS Service .... blog posts reviews some best practices for combating **BGP leaks** and hijacks.

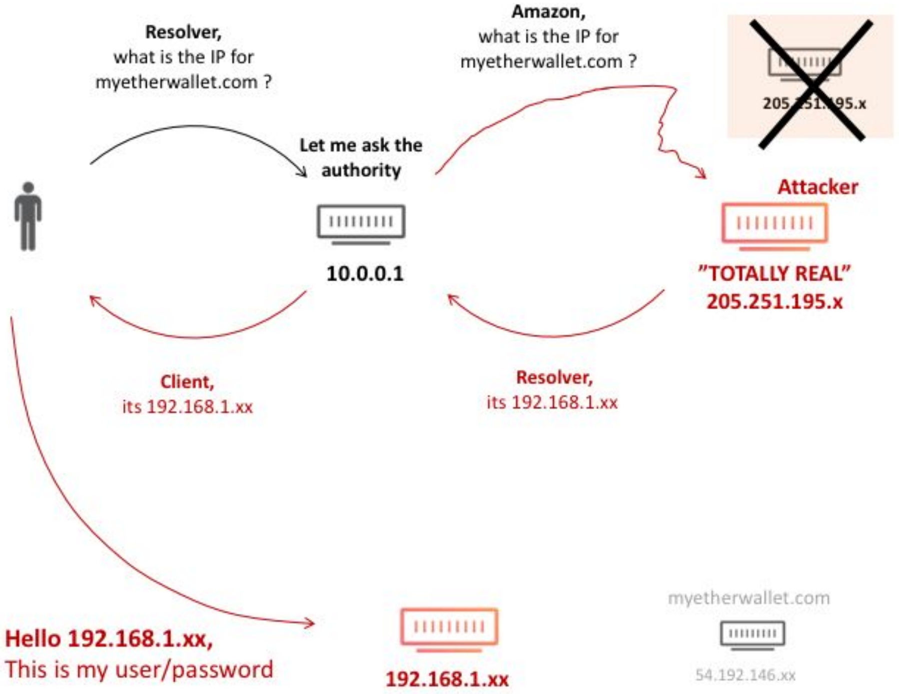
BGP routing security flaw caused Amazon Route 53 incident
https://searchsecurity.techtarget.com/.../BGP-routing-security-flaw-caused-Amazon-Ro...
Apr 25, 2018 - A long-standing flaw in **BGP** routing security that allows attackers to ... to eliminate **BGP** route hijacking, route **leaks** and forwarding of traffic with ...

Suspicious Event Hijacks Amazon Traffic For 2 hours, Steals ...
https://it.slashdot.org/.../suspicious-event-hijacks-amazon-traffic-for-2-hours-steals-cry...
Apr 24, 2018 - **Amazon** lost control of some of its widely used cloud services for two ... 'Kernel Memory **Leaking**' Intel Processor Design Flaw Forces Linux, Windows Redesign ..... We have yet to see a **BGP** session be hijacked, or an external ...

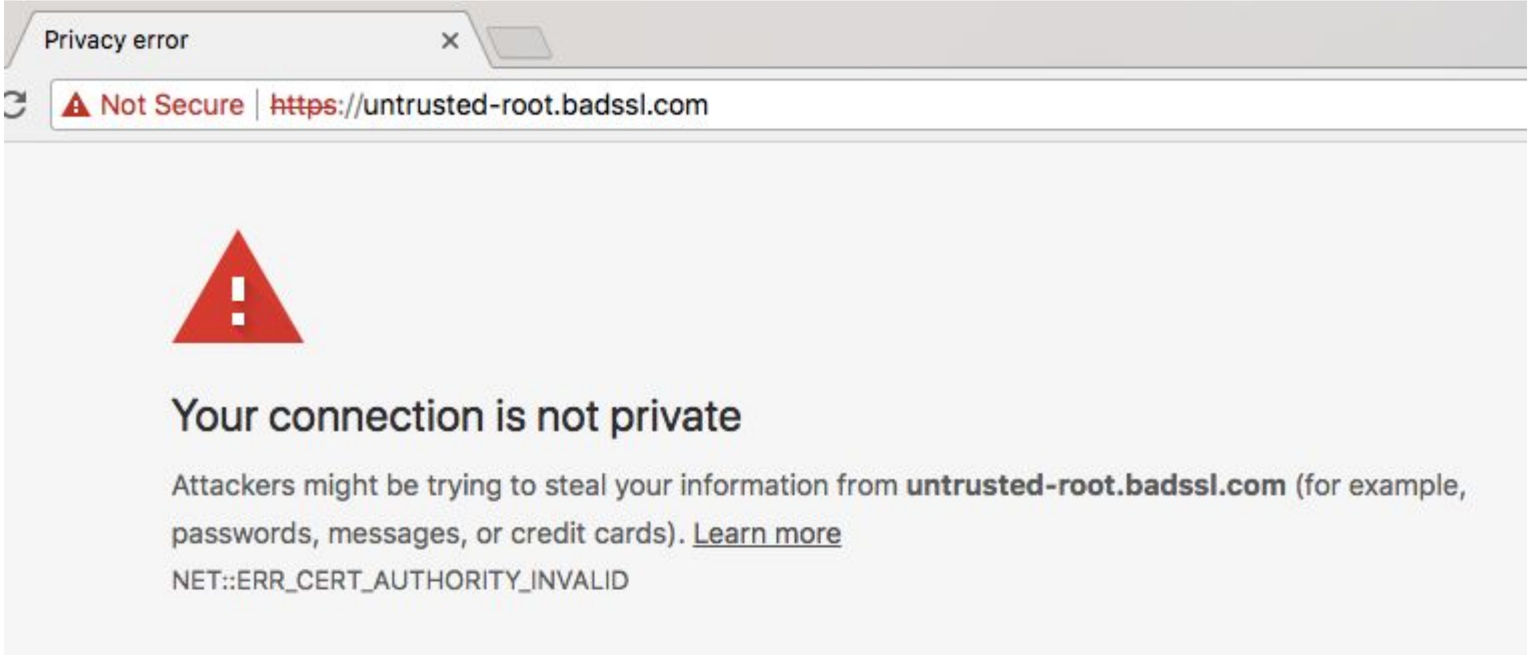
What happened?



What happened?



What happened?



Privacy error

Not Secure | ~~https~~://untrusted-root.badssl.com

Your connection is not private

Attackers might be trying to steal your information from **untrusted-root.badssl.com** (for example, passwords, messages, or credit cards). Learn more

NET::ERR_CERT_AUTHORITY_INVALID

BGP leaks/hijacks

“CIA Triad”: *Confidentiality, integrity, availability*

- Rendering a ressource unreachable (availability)

or

- Impersonating
  - Protocols at risk: DNS/UDP due to no confidentiality nor integrity checks
  - HTTPS and DNSSEC offers a layer of security: reduce availability in exchange of integrity

BGP leaks/hijacks

Someone controlling **65002** wants to hijack **2001:db8:3000::/32** originally announced by **65001**
Possible types:

#	Announcement	AS Path	Effect
1	2001:db8:3000::/32	AS65002	May become shortest AS Path. BGP origin validation/RPKI could filter it out. Sensitive on IXes.
2	2001:db8:3000::/32	AS65002 AS65001	BGP origin validation out of scope. But AS Path longer so less risks. Sensitive on IXes.
3	2001:db8:3000::/48	either	Most specific prefix: will be preferred as long as accepted. BGP origin validation/RPKI could filter it out.

BGP leaks/hijacks

From the previous table: very localized attacks.

While waiting on RPKI:

- IRR filtering: but no guarantees the owner of the prefix actually wrote the information.
- Announcing max-length /24 IPv4 or /48 IPv6 for critical ressources like authority DNS

At scale

At scale

What does “at scale” mean?

- Addresses in all 5 regions (LACNIC, Afrinic, APNIC, ARIN, RIPE)
- Automate prefixes signing and invalidation + long term maintenance
- Strict validation at scale
- Monitoring and failure models

Choice of mode

Hosted or delegated?

- **Hosted**: the certificate and ROA signing is maintained by the RIR
- **Delegated**: a certificate indicates the location of the PKI of the organization. ROAs are generated and signed by the organization.

RIR	Status
Afrinic	Both
APNIC	Both
ARIN	Both
LACNIC	Hosted
RIPE	Hosted and on-demand delegated

Hosted

We chose **hosted** because:

- We do not allocate IP addresses
    - Very few changes, made by the network team

- Only a handful of software for maintaining RPKI CA
    - rsync to maintain

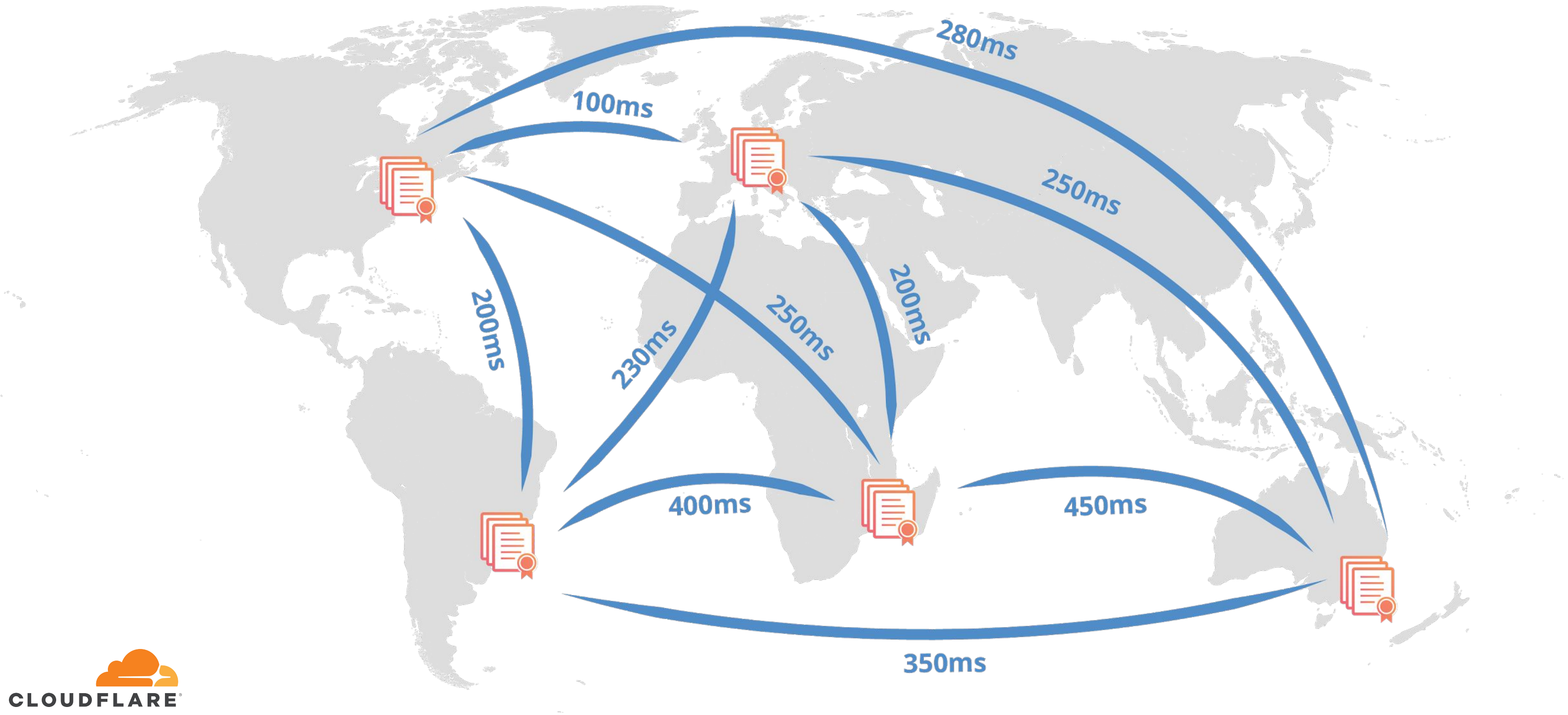
- Not all RIR offer delegated
- If the RIR certificate is compromised: similar to any CA compromised

APIs

- With automation, we want **APIs (GET/PUT/UPDATE/DELETE)**.
- Cloudflare announces many prefixes. We have our provisioning databases/IPAM.

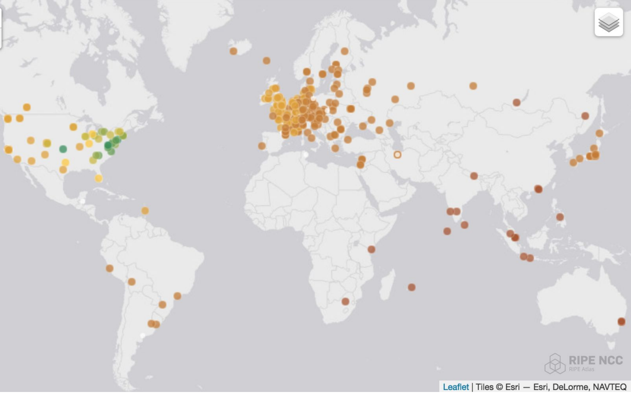
RIR	**Status**
Afrinic	Uses APNIC software
APNIC	Draft
ARIN	Insertion only (not listing, updating, deleting)
LACNIC	No (but easier to batch)
RIPE	No (but easier to batch)

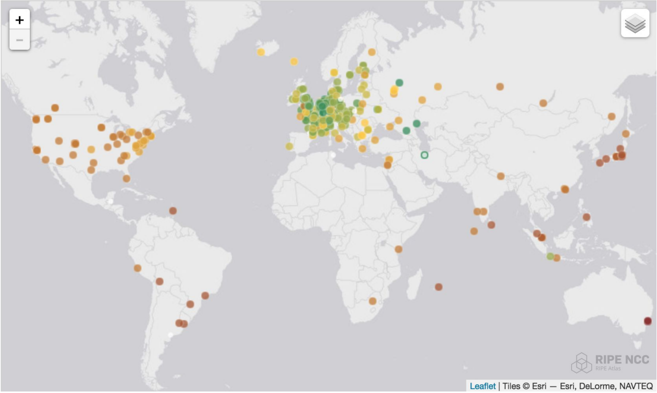
Availability

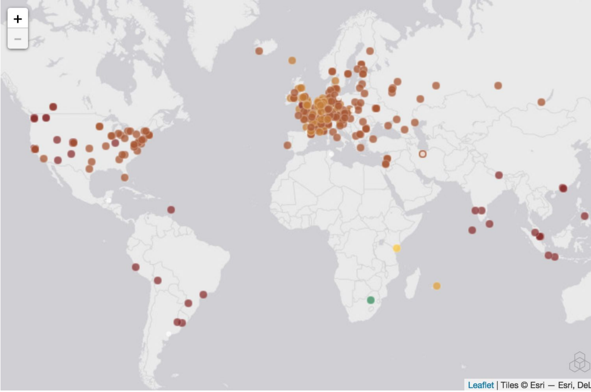


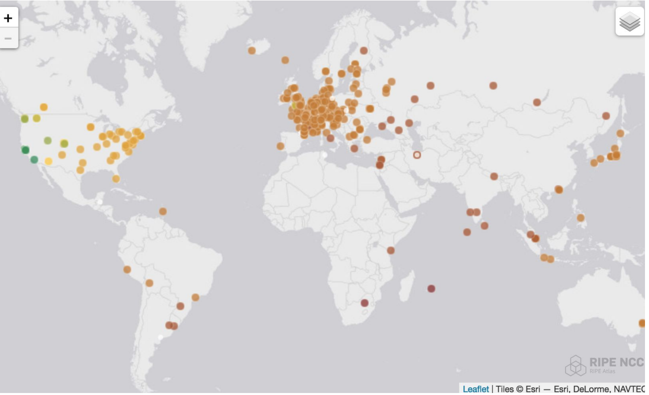
CLOUDFLARE

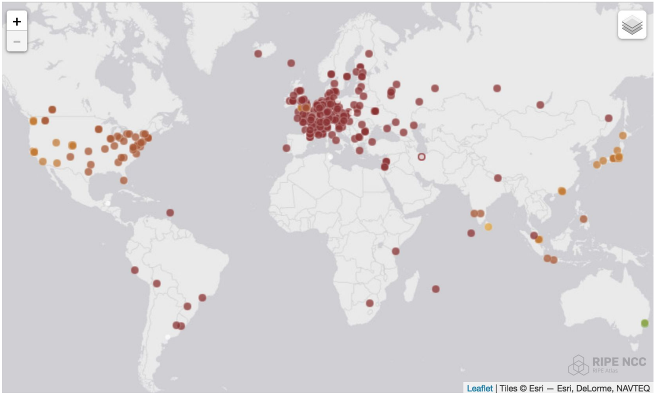
RIPE Atlas → RPKIs

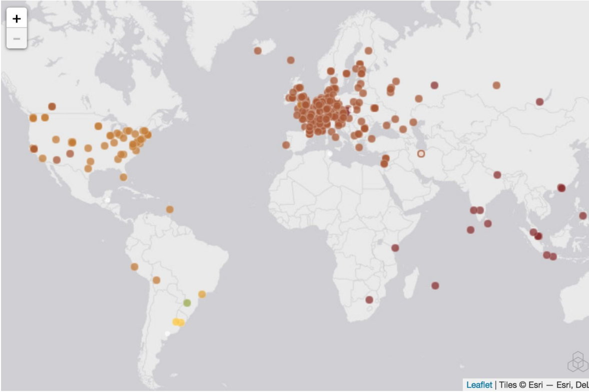












Availability

Low throughput for ⅘ RPKIs (>80ms)

East Asia = no local RPKI

Rsync protocol

- Caching?
- High usage?

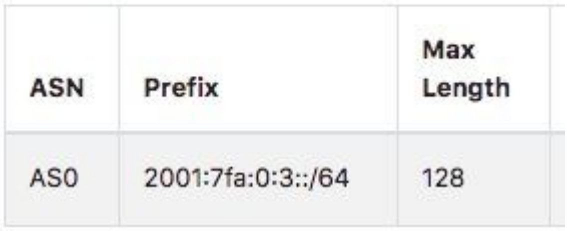
Availability

From Sydney

- RIPE: 90MB, took 5 min (2.4Mbps)
- ARIN: 9MB, took 5 seconds (14Mbps)
- APNIC: 5MB, took 1 second (40Mbps)
- LACNIC: 19MB, took 10 seconds (15Mbps)
- Afrinic: 2MB, took 11 seconds (1.45Mbps)

Future?

- A bit more than 10% of the routes.
- If everything was signed, 1 GB to download at 2-4Mbps (30mn-1 hour)
  - Painful updates/refresh
  - Database could be filled with random records



ASN	Prefix	Max Length
AS0	2001:7fa:0:3::/64	128

Security and performance

We have 150+ PoPs.

How to do validation on every single one of them?

Security and performance

- RTR from central point to each router
  - Single point of failure
  - Latency/packet loss
  - **No encryption** (only TCP supported, no TLS or SSH)

Security and performance

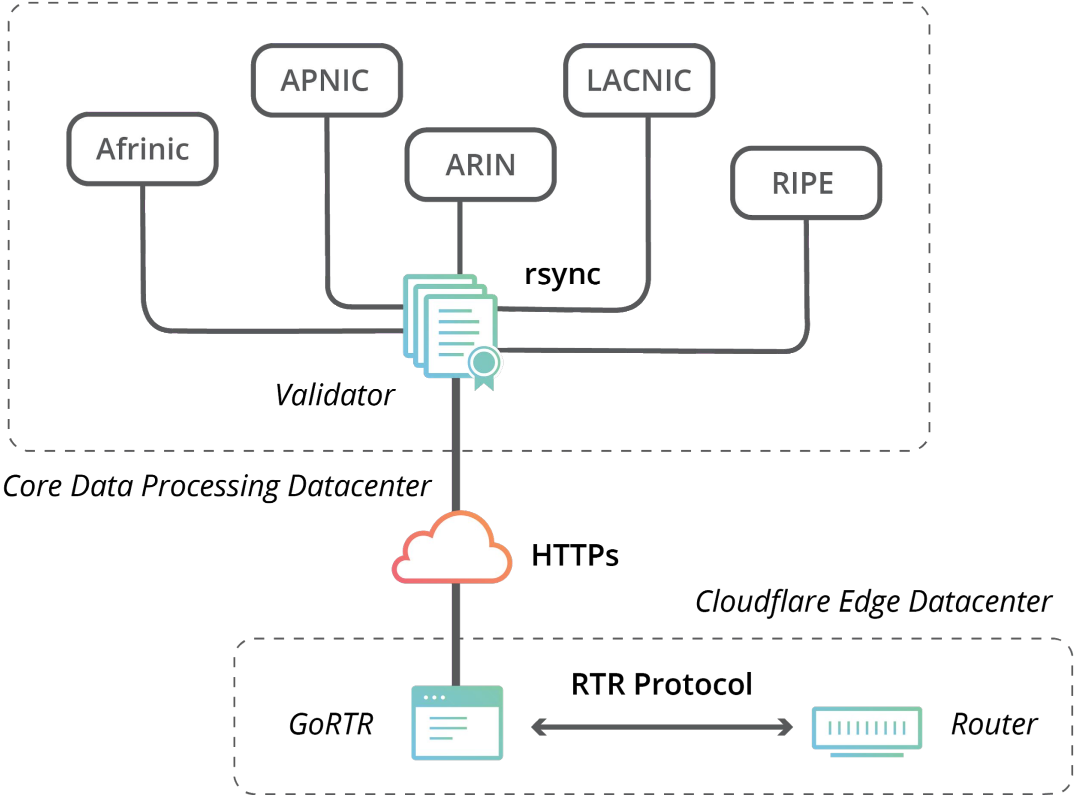
- Validator software on every PoP
    - Wasted resources (10GB disk/RAM, 1-2 CPU)
    - Harder monitoring and maintenance
    - Latency to rsync from faraway places

Security and performance

- Our solution
    - Have a local cache in each PoP using our CDN and HTTPs
    - Central validation having authority
    - Custom RTR software to communicate with routers

    - Integration with Salt and our pipelines

Security and performance



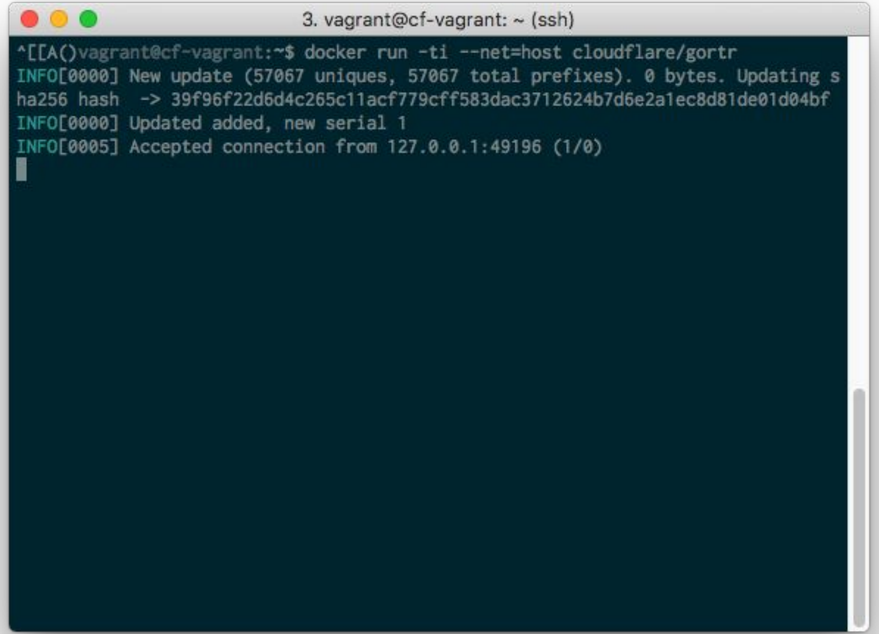
CLOUDFLARE

Also

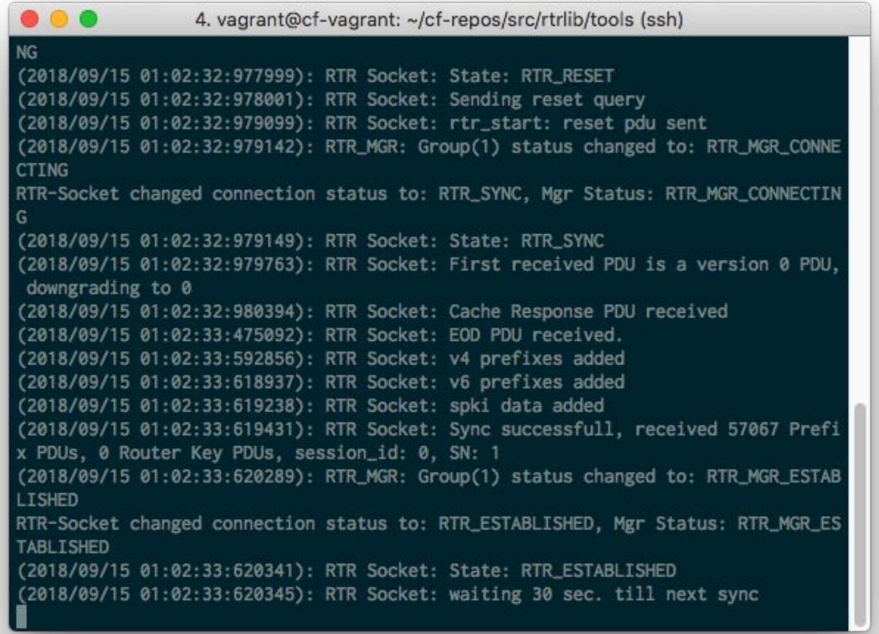
- Use our list of validated prefixes **signed** (RIPE Validator Format):
    - **https://rpki.cloudflare.com/rpki.json**

- Use our implementation of RTR Cache
    - **https://github.com/cloudflare/gortr**

GoRTR



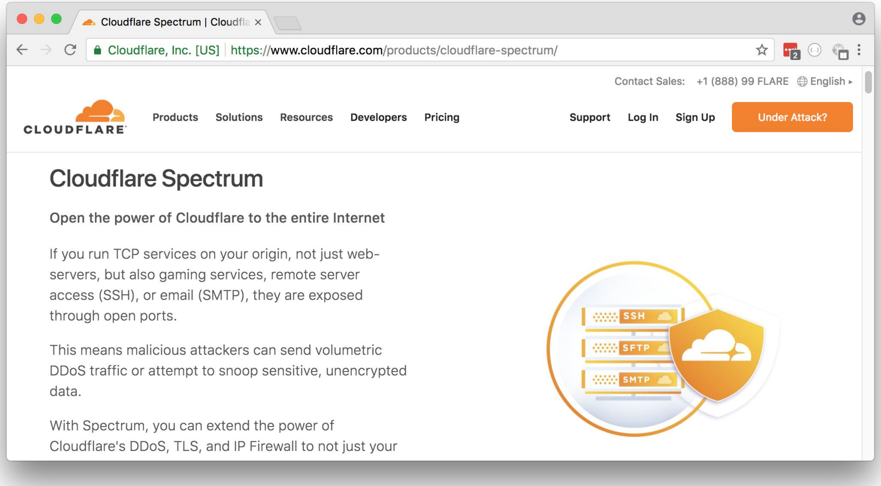
```
^[[A()vagrant@cf-vagrant:~$ docker run -ti --net=host cloudflare/gortr
INFO[0000] New update (57067 uniques, 57067 total prefixes). 0 bytes. Updating s
ha256 hash  -> 39f96f22d6d4c265c11acf779cff583dac3712624b7d6e2a1ec8d81de01d04bf
INFO[0000] Updated added, new serial 1
INFO[0005] Accepted connection from 127.0.0.1:49196 (1/0)
```



```
NG
(2018/09/15 01:02:32:977999): RTR Socket: State: RTR_RESET
(2018/09/15 01:02:32:978001): RTR Socket: Sending reset query
(2018/09/15 01:02:32:979099): RTR Socket: rtr_start: reset pdu sent
(2018/09/15 01:02:32:979142): RTR_MGR: Group(1) status changed to: RTR_MGR_CONNE
CTING
RTR-Socket changed connection status to: RTR_SYNC, Mgr Status: RTR_MGR_CONNECTIN
G
(2018/09/15 01:02:32:979149): RTR Socket: State: RTR_SYNC
(2018/09/15 01:02:32:979763): RTR Socket: First received PDU is a version 0 PDU,
 downgrading to 0
(2018/09/15 01:02:32:980394): RTR Socket: Cache Response PDU received
(2018/09/15 01:02:33:475092): RTR Socket: EOD PDU received.
(2018/09/15 01:02:33:592856): RTR Socket: v4 prefixes added
(2018/09/15 01:02:33:618937): RTR Socket: v6 prefixes added
(2018/09/15 01:02:33:619238): RTR Socket: spki data added
(2018/09/15 01:02:33:619431): RTR Socket: Sync successfull, received 57067 Prefi
x PDUs, 0 Router Key PDUs, session_id: 0, SN: 1
(2018/09/15 01:02:33:620289): RTR_MGR: Group(1) status changed to: RTR_MGR_ESTAB
LISHED
RTR-Socket changed connection status to: RTR_ESTABLISHED, Mgr Status: RTR_MGR_ES
TABLISHED
(2018/09/15 01:02:33:620341): RTR Socket: State: RTR_ESTABLISHED
(2018/09/15 01:02:33:620345): RTR Socket: waiting 30 sec. till next sync
```

Also

- Soon™:
  - A RTR Server service on Cloudflare Spectrum
  - Nothing to install
  - *If you want to run tests*



Cloudflare Spectrum | Cloudfla ×
Cloudflare, Inc. [US] | https://www.cloudflare.com/products/cloudflare-spectrum/
Contact Sales: +1 (888) 99 FLARE
English ▸
CLOUDFLARE
Products
Solutions
Resources
Developers
Pricing
Support
Log In
Sign Up
Under Attack?
SSH
SFTP
SMTP

Cloudflare Spectrum

Open the power of Cloudflare to the entire Internet

If you run TCP services on your origin, not just web-servers, but also gaming services, remote server access (SSH), or email (SMTP), they are exposed through open ports.

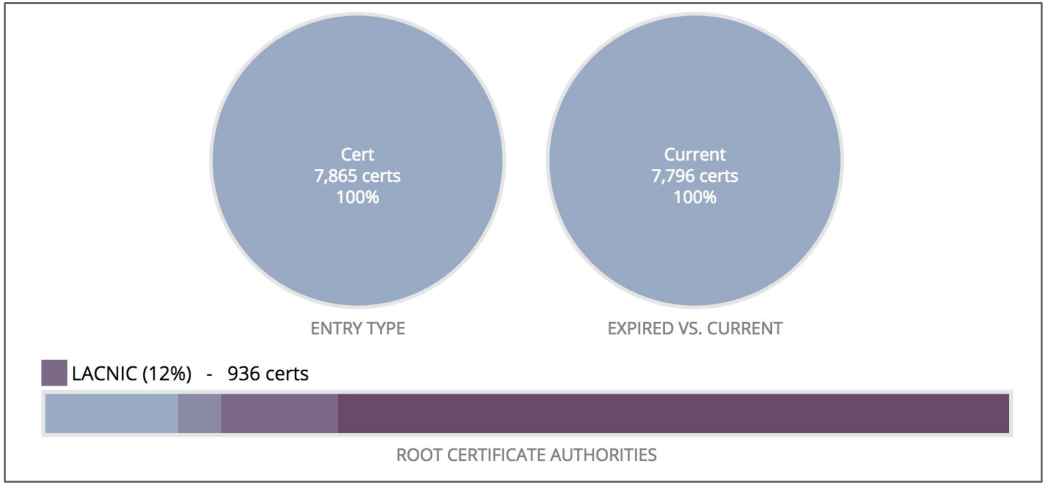
This means malicious attackers can send volumetric DDoS traffic or attempt to snoop sensitive, unencrypted data.

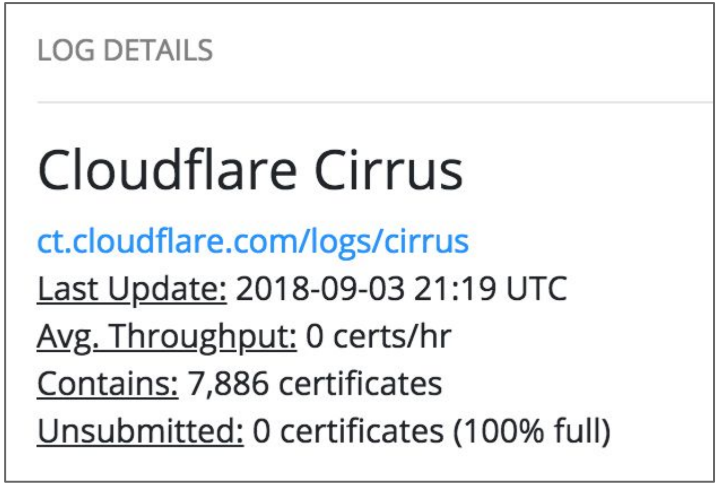
With Spectrum, you can extend the power of Cloudflare's DDoS, TLS, and IP Firewall to not just your

Monitoring

Monitoring of PKI

- Cloudflare’s Certificate Transparency
  - https://ct.cloudflare.com/logs/cirrus





LOG DETAILS

Cloudflare Cirrus

ct.cloudflare.com/logs/cirrus

Last Update: 2018-09-03 21:19 UTC

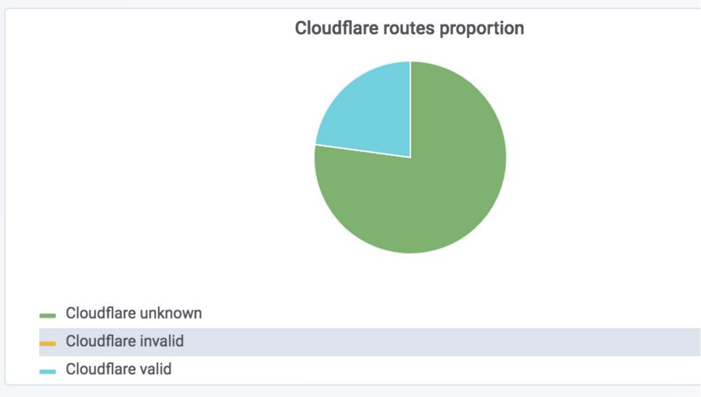
Avg. Throughput: 0 certs/hr

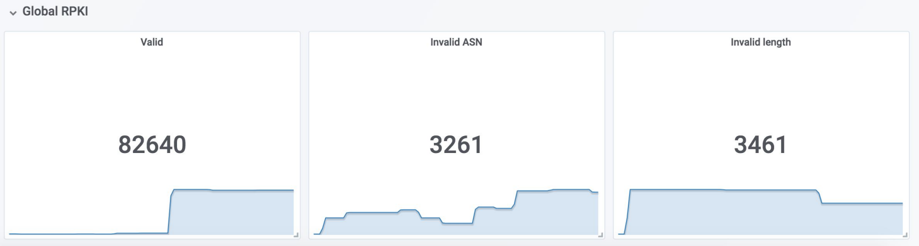
Contains: 7,886 certificates

Unsubmitted: 0 certificates (100% full)

Monitoring of validation

- Coming from our validator:
  - Number of ROAs
  - Distribution
- Coming from our edge
  - Number of invalids/valids
  - Number of filtered routes
- Online
  - https://rpki-monitor.antd.nist.gov/





Monitoring of filtering

- Project @ Cloudflare:
  - With Cloudflare’s presence in more than 180 IX
  - Announce a prefix /24 IPv4 and /48 IPv6 which should be invalid
  - Have the enclosing prefix announced somewhere.
  - Probe the equipments + prefixes announced

Questions?

Thank you!

louis@cloudflare.com
@lpoinsig